

Title:

A Poor Person's Posterior Predictive Checking of Structural Equation Models

Authors:

Taehun Lee

Li Cai

Megan Kuhfeld

Journal publication date:

2016

Published in:

*Structural Equation Modeling*, 23(2), 206-220

IES grant information:

Grant number R305D140046

Funded by National Center for Education Research (NCER)

January 18, 2015, *Structural Equation Modeling*

## A Poor Person's Posterior Predictive Checking of Structural Equation Models

TAEHUN LEE  
UNIVERSITY OF OKLAHOMA, NORMAN

LI CAI  
MEGAN KUHFELD  
UNIVERSITY OF CALIFORNIA, LOS ANGELES

This research was supported in part by an Institute of Education Sciences statistical methodology grant (R305D140046) and a grant from the Research Council of the University of Oklahoma Norman Campus. The views expressed here belong to the authors and do not reflect the views or policies of the funding agencies or grantees. The authors wish to thank Michael Seltzer for insightful comments.

Address all correspondence to: Li Cai, CRESST, UCLA, Los Angeles, CA, USA 90095-1521. Email: lcai@ucla.edu. Phone: 310.794.7136. Fax: 310.825.3883.

## A POOR PERSON'S POSTERIOR PREDICTIVE CHECKING OF STRUCTURAL EQUATION MODELS

**Abstract**

Posterior Predictive Model Checking (PPMC) is a Bayesian model checking method that compares the observed data to (plausible) future observations from the posterior predictive distribution. We propose an alternative to PPMC in the context of structural equation modeling, which we term the Poor Persons PPMC (PP-PPMC), for the situation wherein one cannot afford (or is unwilling) to draw samples from the full posterior. Using only byproducts of likelihood-based estimation (maximum likelihood estimate and information matrix), the PP-PPMC offers a natural method to handle parameter uncertainty in model fit assessment. In particular, a coupling relationship between the classical p-values from the model fit chi-square test and the predictive p-values from the PP-PPMC method is carefully examined, suggesting that PP-PPMC may offer an alternative, principled approach for model fit assessment. We also illustrate the flexibility of the PP-PPMC approach by applying it to case-influence diagnostics.

## 1 Introduction

The posterior predictive model checking (PPMC) method is a Bayesian model diagnostic tool for assessing the compatibility of a posited model to observed data by comparing the observed data to plausible/future observations simulated from the *posterior predictive* distribution. This method is predicated upon the idea that, if the model fits the data reasonably well, the future/plausible observations should be "similar" to the observed data, whereas large discrepancies should be taken as an indication of model misspecification. In applications, "similarity" or "discrepancy" is to be defined (by the researcher) in such a way that differences between a critical aspect of the observed data and that of the model-implied future observations can be properly measured.

The use of model-implied future observations in model-data fit evaluations was first introduced by Guttman (1967), and its formal definition was given by Rubin (1981, 1984), further elaborated by Meng (1994), Gelman, Meng, and Stern (1996). A good didactic discussion of the PPMC method in general can be found in Gelman, Carlin, Stern, and Rubin (2003, chap. 6).

In social and behavioral science research, most of the applications of the PPMC method can be found in the context of the item response theory (IRT) models in assessing item fit, person fit, and dimensionality (Hoijtink & Molenaar, 1997; Janssen, Tuerlinckx, Meulders, & de Boeck, 2000; Glas & Meijer, 2003; Sinharay, Johnson, & Stern, 2006; Levy, Mislevy, & Sinharay, 2009; Levy & Svetina, 2011).

In factor analysis and structural equation modeling (SEM), despite the rapid growth of Bayesian approaches (Arminger & Muthén, 1998; Lee & Zhu, 2000; Ansari, Jedidi, & Dube, 2002; Lee, 2007; Lee, Song, & Tang, 2007; Palomo, Dunson, & Bollen, 2007), most of previous studies focused on Bayesian model building and parameter estimation, paying relatively scant attention to the issue of model fit appraisal and diagnostics. Scheines, Hoijtink, and Boomsma (1999), Levy (2011), and B. Muthén and Asparouhov (2012) are some important exceptions, illustrating key features of Bayesian approaches to model

diagnostics and assessing data-model fit of structural equation models.

In applications of factor analysis and structural equation modeling, standard estimation and model checking methods are based on the method of maximum likelihood (ML). SEM software programs routinely print in the output the ML point estimates for the parameters and the associated standard error estimates. The ML point estimates then replaces the unknown model parameters to yield model-implied mean vectors and covariance matrices from which chi-square model fit test statistics and various practical fit indices such as RMSEA (Root Mean Square Error of Approximation), TLI (Tucker-Lewis Index), or SRMR (Standardized Root Mean Square Residual) can be computed.

In principle, the reference distributions of these test statistics and fit indexes can be developed using asymptotic arguments (Jöreskog, 1993; Ogasawara, 2001) or with the bootstrap (Bollen & Stine, 1993; Efron & Tibshirani, 1994). In the practice of structural equation modeling, however, the asymptotic arguments may not be tenable, despite their general applicability. For instance, the form of a number of popular model fit indices may not be particularly amenable to asymptotic derivations (perhaps with the exception of RMSEA). In addition, the development of asymptotic distribution theory for any new researcher-defined fit index would require advanced statistical prowess. It is unfair to require an average user of structural equation modeling to perform such derivations. On the other hand, while the bootstrap method may sidestep many of the inherent issues with large-sample arguments, it remains computationally demanding because the model under investigation must be fitted to each of the (often hundreds of) bootstrap resamples.

Even setting practicality aside, classical likelihood-based approaches evaluate the discrepancy between the observed data and the hypothesized model when the unknown model parameters are replaced by the *best-fitting* point estimates. The parameter estimation uncertainty, although quantifiable in the form of asymptotic covariance matrix (inverse of Fisher information matrix) of the ML parameter estimates, is not accounted for in classical likelihood-based model fit assessment. This is made explicit in the appli-

cation of the so-called parametric bootstrap method to model fit testing. In parametric bootstrap resampling, the resamples are generated from the hypothesized model with all parameters replaced by the sample ML estimate (see, e.g., Tollenaar & Mooijaart, 2003). The construction of the bootstrapped reference distribution of a fit statistic requires fitting the hypothesized model to each bootstrap resample.

By contrast, the PPMC method is a simulation-based model checking method, requiring neither asymptotic arguments (cf. likelihood-ratio chi-square statistic) nor computationally intensive model re-fitting (cf. bootstrap). When the plausible values of the data can be drawn from the posterior predictive distribution, constructing reference distributions of any test quantity defined by the investigator involves minimal cost. Moreover, use of the posterior predictive distribution implies that one must take into account the entire posterior distribution of the model's parameters, rather than the best-fitting point estimates only. As such, the PPMC method naturally integrates parameter uncertainty into model fit assessment (Gelman et al., 2003; Rupp, Dey, & Zumbo, 2004; Levy, 2011).

To take full advantage of the PPMC method, however, the researcher must be able to simulate draws from the full posterior distribution of the model parameters. If parameter estimation is accomplished with Bayesian sampling based methods, e.g., Markov chain Monte Carlo (MCMC; see e.g., Gilks, Richardson, & Spiegelhalter, 1996), samples from the posterior predictive distribution can be obtained as a byproduct of the MCMC output. On the other hand, for those who chose not to adopt Bayesian methods, are unfamiliar with Bayesian methods, or when Bayesian methods are cumbersome, complexities remain. For example, the choice of prior distribution on variance components can be complicated due to its potentially large effect on subsequent inference (Gelman, 2006). Convergence monitoring of MCMC often requires considerable expertise on the part of the user, and in our view should not be fully automated (see discussions in Cowles & Carlin, 1996 and MacCallum, Edwards, & Cai, 2012).

For high-dimensional highly-parameterized latent variable models, numerous au-

thors advocated that one of the critical first steps toward a sensible full Bayesian analysis in fact lies in the use of a mode-finding method (e.g., ML) for parameter estimation (see e.g., Gelman et al., 2003). Recognizing both the enormous advantages as well as the potential complications of the Bayesian PPMC methods, and at the same time, given the current dominance of likelihood-based methods for parameter estimation and model fit testing (due to both history and practicality), we pose a question that provides the guiding motivation for this research: Can one find a predictive model checking method for evaluating models fitted using the method of maximum likelihood?

In response to this question, we propose a Poor Person's PPMC method (PP-PPMC), which employs byproducts of maximum likelihood estimation, i.e., the ML parameter estimates and the associated asymptotic error covariance matrix. This method is termed a Poor-Person's PPMC because we believe that it may provide a computationally efficient non-iterative (cf. bootstrap) mechanism to conduct predictive model checking that directly builds parameter uncertainty into consideration (cf. standard likelihood ratio test) for the researcher who, for various reasons, cannot "afford to" or is not (yet) willing to draw samples from the full posterior distribution of the model's parameters.

The remainder of this paper is organized as follows. First, we introduce the original PPMC method to provide the necessary background for discussing the proposed PP-PPMC method. Because the PP-PPMC method employs only byproducts of maximum likelihood estimation, we closely examine the relationship between the classical  $p$ -value under the likelihood ratio chi-square test and the  $p$ -values under the PP-PPMC method. In subsequent sections, we apply the PP-PPMC method to two specific cases: overall goodness-of-fit assessment, and case-influence diagnostics. Using empirical and simulated data, we show that the proposed method can be an effective tool for both overall model fit testing and case-influence diagnostics. We conclude the paper with discussions about practical implications of the proposed method and future research.

## 2 Posterior Predictive Model Checking

Let  $\mathcal{E}$  be some sample space. Throughout the paper we use the notation  $\mathbf{Y}^{obs} \in \mathcal{E}$  to represent observed data,  $H$  the hypothesized model, and  $\boldsymbol{\theta}$  the  $d$ -dimensional vector of unknown model parameters that resides in the parameter space  $\boldsymbol{\Theta}$  which is a subset of the  $d$ -dimensional Cartesian product of real numbers  $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$ . We use the notation  $\mathbf{Y}^{rep} \in \mathcal{E}$  to denote hypothetical replicated data or plausible future observations under the hypothesized model  $H$ .

### 2.1 Classical Estimation and Inference for SEM

In this paper we consider a general mean and covariance structure model. Let there be  $p$  observed/manifest variables in the model. The  $p \times 1$  vector of manifest variables  $\mathbf{y}$  serve as indicators for a vector of  $q$  latent variables  $\boldsymbol{\eta}$  via a factor analytic measurement model  $\mathbf{y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}$ , where the unique factors in  $\boldsymbol{\epsilon}$  have zero means and covariance matrix  $\boldsymbol{\Psi}$ . The relationships among the latent variables are described by simultaneous equations  $\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}$ , where the disturbance terms in  $\boldsymbol{\zeta}$  have zero means and covariance matrix  $\boldsymbol{\Phi}$ . Both  $\boldsymbol{\Psi}$  and  $\boldsymbol{\Phi}$  are functions of the parameter vector  $\boldsymbol{\theta}$ . In addition, the measurement intercepts  $\boldsymbol{\tau}$ , the factor loadings  $\boldsymbol{\Lambda}$ , the latent regression intercepts  $\boldsymbol{\alpha}$  and the regression coefficients  $\mathbf{B}$  are also functions of  $\boldsymbol{\theta}$ . Assuming orthogonality of  $\boldsymbol{\zeta}$  and  $\boldsymbol{\epsilon}$ , the following mean and covariance structure model for random vector  $\mathbf{y}$  can be derived:

$$E(\mathbf{y}) = \boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\tau} + \boldsymbol{\Lambda}\mathbf{B}_*^{-1}\boldsymbol{\alpha}, \quad (1)$$

$$cov(\mathbf{y}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\mathbf{B}_*^{-1}\boldsymbol{\Phi}(\mathbf{B}_*^{-1})'\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \quad (2)$$

where  $\mathbf{B}_* = \mathbf{I}_q - \mathbf{B}$  is assumed to be non-singular, and  $\mathbf{I}_q$  is a  $q \times q$  identity matrix.

Let  $p_* = p + p(p+1)/2$  denote the number of observed first and second moments. It is often convenient to consider the  $p_* \times 1$  vector of unique model-implied means and covariances  $\boldsymbol{\sigma}(\boldsymbol{\theta}) = [\boldsymbol{\mu}(\boldsymbol{\theta})', vech(\boldsymbol{\Sigma}(\boldsymbol{\theta}))']'$ , where  $vech(\boldsymbol{\Sigma})$  is the half-vectorization operator that returns the a vector consisting of the  $p(p+1)/2$  unique elements of  $\boldsymbol{\Sigma}$ .

Specification of a structural equation model  $H$  includes the pattern and values of free and fixed elements of the parameter matrices as well as additional restrictions on  $\theta$ . In the classical model fitting context (see, e.g., Cudeck & Henly, 1991), it is often assumed that there exists a true population mean vector  $\mu_0$  and a true population covariance matrix  $\Sigma_0$ , and the model is referred to as correctly specified if and only if there exists some  $\theta_0 \in \theta$  such that  $\sigma(\theta_0) = \sigma_0 = [\mu'_0, \text{vech}(\Sigma_0)']'$ . We assume that the model is locally identified such that the  $p_* \times d$  Jacobian matrix

$$\Delta(\theta) = \frac{\partial \sigma(\theta)}{\partial \theta'} \quad (3)$$

has full column rank, at least in a neighborhood of  $\theta_0$ .

Assuming correct model specification, the model fitting task is reduced to that of parameter estimation. Given a random sample  $\mathbf{Y}^{obs}$  of size  $N$ , the sample mean vector can be written as  $\bar{\mathbf{y}} = N^{-1} \sum_i^N \mathbf{y}_i^{obs}$ , where  $\mathbf{y}_i^{obs}$  denotes the  $i$ th sample observation, and the sample covariance matrix (maximum likelihood estimate)<sup>1</sup> is denoted as  $\mathbf{S} = N^{-1} \sum_{i=1}^N (\mathbf{y}_i^{obs} - \bar{\mathbf{y}})(\mathbf{y}_i^{obs} - \bar{\mathbf{y}})'$ . The  $p_* \times 1$  sample counterpart to  $\sigma(\theta)$  is  $\mathbf{s} = [\bar{\mathbf{y}}', \text{vech}(\mathbf{S})']'$ .

Parameter estimation by the method of normal theory maximum likelihood requires the assumption of multivariate normality of the observed variables. After some algebra, the log-likelihood function can be written as

$$\ell(\theta | \mathbf{Y}^{obs}) = \ell(\theta | \mathbf{s}) \propto -\frac{N}{2} \left\{ \log |\Sigma(\theta)| + \text{tr}[\Sigma(\theta)^{-1} \mathbf{S}] + [\bar{\mathbf{y}} - \mu(\theta)]' \Sigma(\theta)^{-1} [\bar{\mathbf{y}} - \mu(\theta)] \right\}$$

Under multivariate normality, the observed first and second moments in  $\mathbf{s}$  are sufficient statistics under model  $H$ . Maximization of  $\ell(\theta | \mathbf{Y}^{obs})$  with respect to  $\theta$  results in the ML estimate  $\hat{\theta} \in \theta$ . Equivalently, one may also choose to minimize the following maximum

---

<sup>1</sup>For simplicity we do not use the unbiased sample covariance matrix estimate with  $(N - 1)$  as the divisor. All discussions assume large  $N$  so any difference will be negligible.

Wishart likelihood (MWL) fit function:

$$\begin{aligned} T_{\text{MWL}}(\mathbf{Y}^{obs}, \boldsymbol{\theta}) &= [\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta})]' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} [\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta})] \\ &\quad + \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \log |\mathbf{S}| + \text{tr}[\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{S}] - p. \end{aligned} \quad (4)$$

Let  $\mathcal{I}(\boldsymbol{\theta})$  be the (observed) information matrix. It is equal to a half times the second derivative matrix of the MWL fit function.

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{1}{2} \frac{\partial^2 T_{\text{MWL}}(\mathbf{Y}^{obs}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

By reparameterization, we see that

$$N\mathcal{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta} | \mathbf{Y}^{obs})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -\frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial^2 \ell(\boldsymbol{\theta} | \mathbf{Y}^{obs})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'} \frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = N \{ \boldsymbol{\Delta}(\boldsymbol{\theta})' \boldsymbol{\Gamma}(\boldsymbol{\sigma}) \boldsymbol{\Delta}(\boldsymbol{\theta}) \}, \quad (5)$$

where the  $\boldsymbol{\Gamma}(\boldsymbol{\sigma})$  is the information matrix of the (unstructured) mean and covariance parameters in  $\boldsymbol{\sigma}$  based on a multivariate normal distribution,

$$\boldsymbol{\Gamma}(\boldsymbol{\sigma}) = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} \\ & 2^{-1} \mathbf{D}_p' (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) \mathbf{D}_p \end{pmatrix}, \quad (6)$$

and  $\mathbf{D}_p$  is a  $p^2 \times p(p+1)/2$  duplication matrix (see Schott, 2005). The asymptotic distribution of  $\sqrt{N}(\mathbf{s} - \boldsymbol{\sigma}_0)$  is normal (see Browne, 1984),

$$\sqrt{N}(\mathbf{s} - \boldsymbol{\sigma}_0) \xrightarrow{a} \mathcal{N}_{p_*}(\mathbf{0}, \boldsymbol{\Gamma}_0^{-1}), \quad (7)$$

under multivariate normality of the observed variables, where the symbol  $\xrightarrow{a}$  stands for “asymptotically (in  $N$ ) distributed as,” and  $\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}(\boldsymbol{\sigma}_0)$ .

The asymptotic distribution of the ML estimator  $\hat{\theta}$  is given by

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}_d(\mathbf{0}, \mathbf{V}_0), \quad (8)$$

(see Yuan & Bentler, 2007), where the right hand denotes a multivariate normal distribution with zero means and asymptotic covariance matrix equal to the inverse of the information matrix  $\mathbf{V}_0 = \mathcal{I}(\theta_0)^{-1}$ . A consistent finite-sample variance estimate is

$$\hat{\mathbf{V}} = \{N\mathcal{I}(\hat{\theta})\}^{-1} = \{N\Delta(\hat{\theta})'\hat{\Gamma}\Delta(\hat{\theta})\}^{-1}, \quad (9)$$

where  $\hat{\Gamma} = \Gamma(\mathbf{s})$ . In other words the ML estimates are asymptotically normally distributed with large-sample error covariance matrix  $\hat{\mathbf{V}}$ .

The principle of Generalized Least Squares (GLS) leads to another widely used fit function for parameter estimation and statistical inference:

$$T_{\text{GLS}}(\mathbf{Y}^{obs}, \theta) = \frac{1}{2}[\mathbf{s} - \sigma(\theta)]'\Gamma(\mathbf{s})[\mathbf{s} - \sigma(\theta)], \quad (10)$$

Under normality, the GLS and the MWL fit functions lead to asymptotically equivalent solutions (Browne, 1974). It is also well known (see Browne & Arminger, 1995) that under the multivariate normal sampling model for  $\mathbf{y}$ ,  $N$  times the minimized fit function value ( $NT_{\text{MWL}}(\mathbf{Y}^{obs}, \hat{\theta})$  or  $NT_{\text{GLS}}(\mathbf{Y}^{obs}, \hat{\theta})$ ) is distributed as a central chi-square variable with  $p_* - d$  degrees-of-freedom under the null hypothesis of correct model specification when  $N$  tends to infinity.

## 2.2 Basic Setup for PPMC

As mentioned earlier, the essence of PPMC is to compare the observed data  $\mathbf{Y}^{obs}$  with the replicated data  $\mathbf{Y}^{rep}$ , as opposed to focusing on large sample tests. The hypothetical replicated data can be simulated from the posterior predictive distribution, namely the conditional distribution of the replicated data  $\mathbf{Y}^{rep}$  given the observed data  $\mathbf{Y}^{obs}$  and

the posited model  $H$ , denoted as  $p(\mathbf{Y}^{rep}|\mathbf{Y}^{obs}, H)$ . The general form of the posterior predictive distribution is given as

$$p(\mathbf{Y}^{rep}|\mathbf{Y}^{obs}, H) = \int_{\boldsymbol{\theta}} p(\mathbf{Y}^{rep}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|\mathbf{Y}^{obs}, H)d\boldsymbol{\theta} \quad (11)$$

Equation (11) shows that the integral defining the posterior predictive distribution consists of two components. One is  $p(\mathbf{Y}^{rep}|\boldsymbol{\theta}, H)$ , or the sampling distribution of the replicated data  $\mathbf{Y}^{rep}$  given particular values of the model parameters  $\boldsymbol{\theta}$  under model  $H$ . The other is  $p(\boldsymbol{\theta}|\mathbf{Y}^{obs}, H)$  or the posterior distribution of model parameters  $\boldsymbol{\theta}$  under model  $H$  given observed data  $\mathbf{Y}^{obs}$ . We note that  $p(\boldsymbol{\theta}|\mathbf{Y}^{obs}, H)$  quantifies the plausible values of  $\boldsymbol{\theta}$  after the data have been observed. Therefore, Equation (11) shows explicitly how the use of the posterior predictive distribution addresses the problem of unknown model parameters in making probabilistic statements about the replicated data; it integrates out (averages over) the unknown  $\boldsymbol{\theta}$  in the sampling distribution of  $\mathbf{Y}^{rep}$  over its posterior distribution  $p(\boldsymbol{\theta}|\mathbf{Y}^{obs}, H)$ . As such, the entire posterior distribution of the model's parameters is integrated into the model fit checking procedures.

The inferential principle of PPMC is similar to that of classical model fit testing, i.e. to locate the position of observed data  $\mathbf{Y}^{obs}$  in a reference distribution. If the model fits the data well, the observed data  $\mathbf{Y}^{obs}$  will not stand out in the reference distribution. A key difference between PPMC and classical fit testing lies in choice of the reference distribution. Under PPMC, the posterior predictive distribution given in (11) is used as the reference distribution, whereas, under classical testing, the sampling distribution  $p(\mathbf{Y}|\hat{\boldsymbol{\theta}}, H)$ , with  $\boldsymbol{\theta}$  replaced by the ML estimate  $\hat{\boldsymbol{\theta}}$ , provides the basis of reference distributions for model fit hypothesis tests.

### 2.3 Test Quantities and $p$ -values

In order to measure the degrees and manners in which the observed  $\mathbf{Y}^{obs}$  and replicated data  $\mathbf{Y}^{rep}$  are discrepant, appropriate test quantities should be defined (Gelman et

al., 1996). A test quantity  $T(\mathbf{Y}, \boldsymbol{\theta})$ , or measure of discrepancy, is a function of both the data and the parameters. For example, if the overall fit of the hypothesized structural equation model is to be evaluated, the test quantity may be the MWL fit function defined in Equation (4) or the GLS fit function in Equation (10), but treating both data and parameters as arguments.

In classical overall goodness-of-fit testing, the null hypothesis states that the assumed model holds exactly in the population, i.e.,  $H_0 : \sigma_0 = \sigma(\boldsymbol{\theta}_0)$ , where  $\sigma_0 = [\mu'_0, \text{vech}(\Sigma_0)']'$ . Note that the composite null hypothesis contains unknown parameters. In the classical approach, the unknown parameters are replaced by the ML point estimates  $\hat{\boldsymbol{\theta}}$ . The classical  $p$ -values are defined as  $\Pr(\chi^2_{p_* - d} > NT_{\text{MWL}}(\mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}}))$  or  $\Pr(\chi^2_{p_* - d} > NT_{\text{GLS}}(\mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}}))$ , where the probability is taken over a central chi-square distribution with  $p_* - d$  degrees of freedom. Under the null hypothesis, the only source of random variation comes in the form of multivariate normal sampling of the observed data.

Under the PPMC framework, potential test quantities in Equations (4) and (10) are treated as functions of both data  $\mathbf{Y}$  and parameters  $\boldsymbol{\theta}$ . In other words,  $\boldsymbol{\theta}$  in  $T_{\text{MWL}}(\mathbf{Y}, \boldsymbol{\theta})$  is no longer fixed at the ML estimate  $\hat{\boldsymbol{\theta}}$ , and  $\mathbf{Y}$  is no longer only taken to mean the observed data  $\mathbf{Y}^{obs}$ . When the distribution of the test quantity is based on the joint posterior distribution of the replicated data  $\mathbf{Y}^{rep}$  and  $\boldsymbol{\theta}$ , the distribution of  $T(\mathbf{Y}^{rep}, \boldsymbol{\theta})$  is usually called *predictive* test quantities in the Bayesian literature.

As a direct result of using the parameter posterior  $p(\boldsymbol{\theta} | \mathbf{Y}^{obs}, H)$  to quantify uncertainty in  $\boldsymbol{\theta}$ ,  $T(\mathbf{Y}^{obs}, \boldsymbol{\theta})$ , or in other words, the test quantity with data fixed at the observed values but  $\boldsymbol{\theta}$  randomly sampled from its posterior, is often referred to as *realized* test quantity in the Bayesian literature. Note that variation in  $T(\mathbf{Y}^{obs}, \boldsymbol{\theta})$  comes from posterior variability of the parameters.

Under the PPMC framework, the distribution of the predictive test quantity plays the role of the reference distribution. The distribution of the realized test quantity plays the role of the observed test statistics. Thus test quantities can be considered as generaliza-

tions of classical test statistics.

Bayesian posterior predictive  $p$ -values for the test quantities can be defined as

$$\begin{aligned} \text{Bayesian } p\text{-value} &= \Pr\{T(\mathbf{Y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{Y}^{obs}, \boldsymbol{\theta}) | \mathbf{Y}^{obs}, H\} \\ &= \int_{\mathcal{E}} \int_{\boldsymbol{\theta}} \mathbf{1}\left\{T(\mathbf{Y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{Y}^{obs}, \boldsymbol{\theta})\right\} p(\mathbf{Y}^{rep} | \boldsymbol{\theta}, H) p(\boldsymbol{\theta} | \mathbf{Y}^{obs}, H) d\boldsymbol{\theta} d\mathbf{Y}^{rep}, \end{aligned} \quad (12)$$

where  $\mathbf{1}\{T(\mathbf{Y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{Y}^{obs}, \boldsymbol{\theta})\}$  is an indicator function that takes on a value of 1 if and only if the event  $\{T(\mathbf{Y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{Y}^{obs}, \boldsymbol{\theta})\}$  is true. That is, the Bayesian  $p$ -value for a given test quantity  $T(\mathbf{Y}, \boldsymbol{\theta})$  is defined as the probability that the replicated data,  $\mathbf{Y}^{rep}$ , are more extreme than the observed data,  $\mathbf{Y}^{obs}$ , as measured by the test quantity, where the probability is taken over the joint posterior distribution of  $\mathbf{Y}^{rep}$  and  $\boldsymbol{\theta}$ .

Bayesian  $p$ -values may be employed in the same manner as classical  $p$ -values, for rejecting the null hypothesis if the value is less than the nominal significance level  $\alpha$ . From this hypothesis-testing perspective, however, the Type I error rates of the Bayesian  $p$ -value are known to be below the nominal  $\alpha$  level resulting in conservative inferences. Theoretical properties of the Bayesian  $p$ -values are thoroughly examined in Robins, van der Vaart, and Ventura (2000) and Dahl (2006).

Another perspective regarding the use of the Bayesian  $p$ -values, however, is that the Bayesian  $p$ -value provides a simple numerical summary of the degree of discrepancy between the reality and the model, rather than a rigid accept-rejection decision rule. This perspective is based on a widely accepted fundamental principle that all statistical models are wrong to some degree (Box, 1979), and thus the more pertinent issue is to characterize the ways in which the assumed model is wrong and what aspects of the model is useful for description, prediction, and synthesis (Cudeck & Henly, 1991, p. 512). From this perspective, graphical model checking has been advocated (Stern, 2000; Gelman et al., 1996; Gelman, 2003, 2004, 2007), with Bayesian  $p$ -value serving as a numerical summary of the model-data discrepancy.

## 2.4 Implementation

As shown in Equations (11) and (12), implementing the PPMC method involves multidimensional integration, which in most circumstances is analytically intractable. To circumvent analytical derivations of the posterior predictive distribution, Monte Carlo simulation methods are employed in practice. Specifically, a composition method is used. Note that by Equation (11), the joint posterior of  $\mathbf{Y}^{rep}$  and  $\boldsymbol{\theta}$  can be obtained as a product:  $p(\mathbf{Y}^{rep}, \boldsymbol{\theta} | \mathbf{Y}^{obs}, H) = p(\mathbf{Y}^{rep} | \boldsymbol{\theta}, H)p(\boldsymbol{\theta} | \mathbf{Y}^{obs}, H)$ .

First,  $L$  sets of plausible parameters  $\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^L$  are sampled from the posterior distribution of the parameters  $p(\boldsymbol{\theta} | \mathbf{Y}^{obs}, H)$ . For each plausible parameter vector  $\boldsymbol{\theta}^\ell$ , we draw one hypothetical replicated data set  $\mathbf{Y}^{rep,\ell}$  from the sampling distribution  $p(\mathbf{Y} | \boldsymbol{\theta}^\ell, H)$ . We then have  $L$  pairs of draws from the joint posterior distribution of  $\mathbf{Y}^{rep}$  and  $\boldsymbol{\theta}$ , i.e.,

$$(\mathbf{Y}^{rep,\ell}, \boldsymbol{\theta}^\ell) \sim p(\mathbf{Y}^{rep}, \boldsymbol{\theta} | \mathbf{Y}^{obs}, H), \ell = 1, \dots, L. \quad (13)$$

With sufficiently large number of draws, arbitrarily close approximations to functionals of  $p(\mathbf{Y}^{rep}, \boldsymbol{\theta} | \mathbf{Y}^{obs}, H)$  can be constructed with sample averages.

For example, based on these draws, we can compare the distribution of the realized test quantities,  $T(\mathbf{Y}^{obs}, \boldsymbol{\theta}^\ell)$  against that of the predictive test quantities  $T(\mathbf{y}^{rep,\ell}, \boldsymbol{\theta}^\ell)$ . Creating a scatterplot of the predictive test quantities (on the Y-axis) against the realized test quantities (on the X-axis) provides a convenient visual assessment of model adequacy. For correctly specified models, the points are expected to be evenly divided along the 45-degree reference line.

The Bayesian  $p$ -values defined in Equation (12) can be approximated by counting the proportion of draws for which the predictive test quantity exceeds its corresponding realized test quantity. That is,

$$\text{Bayesian } p\text{-value} \approx \frac{1}{L} \sum_{\ell=1}^L \mathbf{1} \left\{ T(\mathbf{Y}^{rep,\ell}, \boldsymbol{\theta}^\ell) \geq T(\mathbf{Y}^{obs}, \boldsymbol{\theta}^\ell) \right\}. \quad (14)$$

This is equivalent to counting the proportion of points above the 45-degree reference line in a scatterplot of the two test quantities defined above.

### 3 Poor-Person's Posterior Predictive Model Checking

Considering the role of the posterior distribution  $p(\theta|\mathbf{Y}^{obs}, H)$  in the construction of the posterior predictive distribution (see Equation 11), it is not too difficult to see that other distributions of the parameters could have been utilized (Gelfand, 1996, p.149). In fact, Box (1980) suggested the use of the *prior* distribution, whereas Bayarri and Berger (2000) proposed the use of the *partial* or *conditional posterior* distribution. The predictive model checking procedures based on these alternatives are now known as the prior predictive model checking and partial (or conditional) posterior predictive model checking method, respectively. It has also been suggested that the *cross-validation posterior* distribution, which is the posterior distribution conditional on a subset of the observed data, can be used for predictive model checking purposes (Geisser, 1975; Gelfand, Dey, & Chang, 1992; Pena & Tiao, 1992). Notice that all of these distributions can be combined into a general expression

$$p(\mathbf{Y}^{rep}|\mathcal{A}, H) = \int_{\theta} p(\mathbf{Y}^{rep}|\theta, H)p(\theta|\mathcal{A}, H)d\theta. \quad (15)$$

For example, with  $\mathcal{A}$  being an empty set, the general expression in (15) becomes the prior predictive distribution, whereas with  $\mathcal{A}$  being the observed data  $\mathbf{Y}^{obs}$ , the general expression in (15) becomes the posterior predictive distribution in (11). With  $\mathcal{A}$  equal to the set denoted by  $\{\mathbf{Y}^{obs}/T(\mathbf{Y}^{obs})\}$ , where  $T(\mathbf{Y}^{obs})$  represents a test statistic summarizing a characteristic of  $\mathbf{Y}^{obs}$  and / is used to indicate “partialing out”  $T(\mathbf{Y}^{obs})$  from  $\mathbf{Y}^{obs}$ , the general expression in (15) becomes the partial posterior predictive distribution. Levy (2011) and others have noted that posterior predictive, prior predictive, and partial predictive model checking can be viewed as structurally similar. We wish to follow this line of reasoning and suggest an alternative that may provide a convenient framework

for predictive model checking.

In this paper, we propose the use of a multivariate normal distribution with its mean vector equal to the ML estimate  $\hat{\boldsymbol{\theta}}$  and dispersion matrix equal to the asymptotic covariance matrix of the the ML estimate  $\hat{\mathbf{V}}$  (see Equation 9). That is, we propose that the distribution  $p(\boldsymbol{\theta}|\mathcal{A}, H)$  in (15) be replaced by the following normal distribution,

$$\varphi_d(\boldsymbol{\theta}) = |2\pi\hat{\mathbf{V}}|^{-1/2} \exp \left\{ -\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \hat{\mathbf{V}}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right\}. \quad (16)$$

This proposal is based on a well known result in the Bayesian literature that asymptotically the contribution of the likelihood tends to dominate in the posterior. As a consequence, the shape of the posterior, to a first approximation, converges to a multivariate normal with its mean centered around the ML estimate and dispersion matrix equal to the inverse of the information matrix (Gelman, 2003).

### 3.1 Some Theoretical Basis

The asymptotic posterior normality can be heuristically shown by expanding the log-likelihood  $\ell(\boldsymbol{\theta}|\mathbf{Y})$  for fixed  $\mathbf{Y}$  around  $\hat{\boldsymbol{\theta}}$  in a multivariate Taylor series, to second order:

$$\ell(\boldsymbol{\theta}|\mathbf{Y}) \approx \ell(\hat{\boldsymbol{\theta}}|\mathbf{Y}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \left[ \frac{\partial^2 \ell(\hat{\boldsymbol{\theta}}|\mathbf{Y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}). \quad (17)$$

Notice that the gradient vector of the log-likelihood vanishes in this expansion because  $\hat{\boldsymbol{\theta}}$  is a stationary point. Given Equation (17), and using the fact that the posterior is proportional to the prior  $p(\boldsymbol{\theta})$  and the likelihood is  $\exp\{\ell(\boldsymbol{\theta}|\mathbf{Y})\}$ , it can be shown that

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{Y}) &\propto p(\boldsymbol{\theta}) \exp\{\ell(\boldsymbol{\theta}|\mathbf{Y})\} \\ &\approx p(\boldsymbol{\theta}) \exp\{\ell(\hat{\boldsymbol{\theta}}|\mathbf{Y}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \left[ \frac{\partial^2 \ell(\hat{\boldsymbol{\theta}}|\mathbf{Y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\} \\ &\propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' [N\mathcal{I}(\hat{\boldsymbol{\theta}})] (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right\} = \exp \left\{ -\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \hat{\mathbf{V}}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right\}. \end{aligned} \quad (18)$$

Recall that  $\hat{\mathbf{V}} = \{N\mathcal{I}(\hat{\boldsymbol{\theta}})\}^{-1}$  (Equation 9). On the final line,  $p(\boldsymbol{\theta}) \exp\{\ell(\hat{\boldsymbol{\theta}}|\mathbf{y})\}$  can be absorbed into the proportionality constant because  $\ell(\hat{\boldsymbol{\theta}}|\mathbf{y})$  is a constant that does not involve  $\boldsymbol{\theta}$  and  $p(\boldsymbol{\theta})$  is assumed to be diffuse (in comparison to the likelihood). Notice that the final expression in Equation (18) represents the kernel of the multivariate normal distribution with mean equal to  $\hat{\boldsymbol{\theta}}$  and the dispersion matrix equal to  $\hat{\mathbf{V}}^{-1}$ , same as Equation (16). More rigorous justifications of posterior normality have been given by many authors, including Le Cam (1953) and Le Cam and Yang (2000).

The asymptotic normality of the posterior implies that the method of maximum likelihood can be regarded as a large-sample Bayesian procedure. This of course requires a conceptual shift from interpreting the parameters as fixed quantities to be uncovered by an estimator such as maximum likelihood to random variables that may have posterior variance conditional on the data and the model. From the multivariate normal approximation to the posterior distribution of  $\boldsymbol{\theta}$ , the posterior predictive distribution, the posterior distributions of test quantities, and the Bayesian  $p$ -values can be constructed. Because we sample from a crude first-order approximation to the posterior, we term the procedure a Poor-Person's posterior predictive model checking (PP-PPMC).

### 3.2 Implementation

To implement PP-PPMC, we first fit the hypothesized model  $H$  to the observed data  $\mathbf{Y}^{obs}$  by minimizing the MWL fit function, which yields the ML estimates of the parameters  $\hat{\boldsymbol{\theta}}$  and the associated asymptotic covariance matrix  $\hat{\mathbf{V}}$ . Second,  $L$  sets of plausible parameter vectors  $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots, \boldsymbol{\theta}^{(L)}\}$  are simulated from the multivariate normal distribution with mean and covariance matrix equal to  $\hat{\boldsymbol{\theta}}$  and  $\hat{\mathbf{V}}$ , respectively. Then for each  $\boldsymbol{\theta}^{(\ell)}$ , a replicated data set  $\mathbf{Y}^{rep,(\ell)}$  is drawn from the sampling distribution under the hypothesized model  $H$ , i.e.,  $p(\mathbf{Y}|\boldsymbol{\theta}^{(\ell)}, H)$ . We now have  $L$  pairs of  $\mathbf{Y}^{rep}$  and  $\boldsymbol{\theta}$  samples, i.e.,  $\{(\mathbf{Y}^{rep,(1)}, \boldsymbol{\theta}^{(1)}), (\mathbf{Y}^{rep,(2)}, \boldsymbol{\theta}^{(2)}), \dots, (\mathbf{Y}^{rep,(L)}, \boldsymbol{\theta}^{(L)})\}$ . Using these  $L$  pairs of  $\mathbf{Y}^{rep}$  and  $\boldsymbol{\theta}$  samples, we can compute  $L$  pairs of predictive and realized values for a given test quantity  $T(\mathbf{Y}, \boldsymbol{\theta})$ , i.e.  $\{T(\mathbf{Y}^{rep,(\ell)}, \boldsymbol{\theta}^{(\ell)}), T(\mathbf{Y}^{obs}, \boldsymbol{\theta}^{(\ell)})\}$  with  $\ell = 1, \dots, L$ . Using these predictive and

realized test quantities, we can create scatterplots and approximate Bayesian  $p$ -values just as outlined for the original PPMC method. We term these predictive  $p$ -values PP-PPMC predictive  $p$ -values:

$$\text{PP-PPMC predictive } p\text{-value} \approx \frac{1}{L} \sum_{\ell=1}^L \mathbf{1} \left\{ T(\mathbf{Y}^{rep,(\ell)}, \boldsymbol{\theta}^{(\ell)}) \geq T(\mathbf{Y}^{obs}, \boldsymbol{\theta}^{(\ell)}) \right\}. \quad (19)$$

Aside from requiring the ability to simulate multivariate normal deviates with given mean vector and covariance matrix, it is clear that the PP-PPMC procedure only requires byproducts of the likelihood-based estimation, i.e.  $\hat{\boldsymbol{\theta}}$  and  $\hat{\mathbf{V}}$ , which are routinely available in output from standard structural equation modeling software programs. The PP-PPMC method can offer a computationally efficient alternative to the original PPMC method, with no need for drawing samples from the full posterior (likely with MCMC). The PP-PPMC can also be an effective alternative to the classical likelihood-ratio chi-square test, offering a way to explicitly account for parameter estimation uncertainty.

#### 4 A Coupling Effect

In the PP-PPMC approach, the use of the normal posterior approximation leads to an interesting coupling relation of the classical  $p$ -values and the PP-PPMC predictive  $p$ -values. The main result can be stated as follows. With the test quantity chosen to be an omnibus overall goodness-of-fit test statistic such as the MWL and GLS fit functions in (4) and (10), the PP-PPMC predictive  $p$ -value can be further approximated as

$$\begin{aligned} \text{Approximate PP-PPMC predictive } p\text{-value} &= \Pr\{T(\mathbf{Y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{Y}^{obs}, \boldsymbol{\theta}) | \mathbf{Y}^{obs}, H\} \\ &\approx \Pr\{\chi_{p_*}^2 - \chi_d^2 > NT(\mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}})\}, \end{aligned} \quad (20)$$

where  $T(\mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}})$  is the minimized value of the fit function based on observed data and  $\chi_{p_*}^2$  and  $\chi_d^2$  represent two independent chi-square random variables with  $p_* = p + p(p+1)/2$  and  $d$  degrees-of-freedom, respectively. A derivation of this result is in the Appendix.

Recall that the classical  $p$ -values can be obtained by

$$\text{classical } p\text{-value} = \Pr\{\chi_{p_*-d}^2 > NT(\mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}})\}. \quad (21)$$

The result in Equation (20) essentially suggests that, when the classical MWL or GLS fit functions are used as test quantities, the PP-PPMC predictive  $p$ -values can be further approximated non-iteratively. Importantly, the same classical model fit test statistics can be used, although a different reference distribution is needed.

In other words, computing the approximate PP-PPMC predictive  $p$ -value requires only the minimum fit function chi-square value and a random number generator that can sample chi-square random variables. For example, given a model with  $p = 9$  manifest variables and  $d = 24$  degrees-of-freedom, if the freely available R programming language is used, the PP-PPMC predictive  $p$ -value associated with a minimum fit function chi-square of 50 can be obtained in as little as 3 lines of code:

```
T <- 50; p <- 9; d <- 24;
dist <- rchisq(10000,df=p*(p+1)/2+p)-rchisq(10000,df=d)
print(mean(ifelse(dist > T, 1, 0)),digit=3)
```

Equations (20) and (21), when taken together, shed light on the relationship between PP-PPMC predictive  $p$ -values and classical  $p$ -values. It is interesting to note that the same minimum fit function value is used for obtaining both the predictive and classical  $p$ -values. Notice that the distribution of the difference  $\chi_{p_*}^2 - \chi_d^2$  has the same mean of  $p_* - d$  as the distribution of  $\chi_{p_*-d}^2$ , but it has a larger variance  $2(p_* + d)$  than that of a  $\chi_{p_*-d}^2$  distribution, being  $2(p_* - d)$ . This implies that the PP-PPMC method, as a result of accounting for parameter estimation uncertainty, tends to yield less stringent evaluations regarding model fit than the classical asymptotic theory based chi-square test.

## 5 Application to Empirical Data

We use the well known Open-Book Closed-Book (OBCB) data set in Mardia, Kent, and Bibby (1979). This data set contains test scores from  $N = 88$  examinees on  $p = 5$  subjects: Mechanics, Vectors, Algebra, Analysis and Statistics from 88 examinees. For this data set, it is well known that a two-factor correlated traits CFA model fits the data extraordinarily well (see e.g., Cai & Lee, 2009). The likelihood ratio chi-square test statistic is equal to 2.07 with 4 degrees-of-freedom and classical  $p$ -value of .72.

We choose the test quantity to be the MWL discrepancy function in Equation (4) for the purpose of the assessment of overall model fit. This choice enables a direct comparison between the PP-PPMC method and the classical likelihood ratio chi-square test. Furthermore, the use of a chi-square type discrepancy function offers opportunities to evaluate the quality of approximation in Equation (20) in empirical data analysis.

Under the PP-PPMC framework, we evaluate the global fit of four hypothesized models including a two-factor CFA model ( $H_1$ ), a congeneric test model ( $H_2$ ), an essentially tau-equivalent model ( $H_3$ ), and a parallel test model ( $H_4$ ) to the OBCB data set. Notice that each of the four models is successively more restrictive than the preceding one.

Specifically, we first obtain the maximum likelihood estimates of the model parameters and the associated asymptotic covariance matrices for each of the four hypothesized models. Then  $\boldsymbol{\theta}^{(\ell)}$  and  $\mathbf{Y}^{rep,(\ell)}$  are simulated from the multivariate normal approximation to joint posterior distribution, with  $p(\boldsymbol{\theta}|\mathbf{Y}^{obs}, H)$  replaced by Equation (16). Based on these simulated draws of  $\boldsymbol{\theta}$  and  $\mathbf{Y}^{rep}$ , predictive and realized test quantities based on (4) and the associated PP-PPMC predictive  $p$ -values are obtained. Finally, the PP-PPMC predictive  $p$ -values are compared to Bayesian PPMC predictive  $p$ -values obtained from Mplus Version 7 (L. K. Muthén & Muthén, 1998-2012).

---

### Insert Figure 1 About Here

Panel (a) in Figure 1 shows the scatterplot of the predictive versus realized test quan-

tities measuring overall discrepancy between the data and the two-factor CFA model ( $H_1$ ) for both the PP-PPMC and PPMC methods. The plus-signs represent the PPMC test quantities and the squares the PP-PPMC test quantities. It is clear that the distributions are largely overlapping. The scattering of the predictive and realized discrepancies is evenly divided along the 45 degree line, yielding an estimated PP-PPMC predictive  $p$ -value of .556. The PPMC predictive  $p$ -value is equal to .577, indicating that the normal approximation in PP-PPMC resulted in a very similar estimate. The other panels (b), (c), and (d), provide the scatterplots comparing the predictive and realized discrepancies for models  $H_2$ ,  $H_3$ , and  $H_4$ , respectively.

In these figures, it is worth noting that the magnitude of realized discrepancies tend to increase, as successively more restrictive (and increasingly mis-specified) models are fitted to data. The PP-PPMC predictive  $p$ -values for the three models are .306, .245, and .000, respectively. Such a decrease in the PP-PPMC predictive  $p$ -values is entirely consistent with our expectations.

Table 1: Model Fit Summary for the Open-Book Closed-Book Data

Model	$NT_{MWL}(\mathbf{Y}^{obs}, \hat{\theta})$	df	$p$ -values			
			Classical	Bayesian	PP-PPMC	Approx. PP-PPMC
$H_1$	2.073	4	0.722	0.577	0.556	0.589
$H_2$	8.978	5	0.110	0.283	0.306	0.308
$H_3$	14.937	9	0.093	0.253	0.245	0.216
$H_4$	79.339	13	0.000	0.000	0.000	0.000

Note.  $T_{MWL}(\mathbf{Y}^{obs}, \hat{\theta})$  represents the minimum MWL fit-function value.

Table 1 shows a summary of PP-PPMC, PPMC, and classical model fit results for the four models. Because the proposed PP-PPMC method employs standard maximum likelihood estimation, the likelihood-ratio test statistics  $NT_{MWL}(\mathbf{Y}^{obs}, \hat{\theta})$  and the associated classical  $p$ -values are readily available. The last two columns show the PP-PPMC predictive  $p$ -values and the corresponding approximate PP-PPMC predictive  $p$ -values, found using the coupling approximation in Equation (20). From the Table we can conclude:

1) the PP-PPMC predictive  $p$ -values are close to the Bayesian PPMC values, 2) the PP-PPMC predictive  $p$ -values decrease as expected for more restrictive models, 3) the quality of coupling approximation is promising, and 4) the PP-PPMC predictive  $p$ -values tend to be less extreme than the corresponding classical  $p$ -values, again as expected.

## 6 Poor-Person's PPMC for Case Diagnostics

The assessment of global model fit via PP-PPMC only tells us the overall extent to which the model is compatible with observed data. When the model fit is poor, more detailed inspections are required to discover the sources of misfit. In this section we illustrate the flexibility of the PP-PPMC method using the diagnosis of *influential cases* as an example. We wish to highlight the fact that other user-defined fit indices or altogether different aspects of model fit assessment can be studied with the same general approach.

The influential cases differ markedly from other cases in the sample in the way they exert influence on model fit. In the context of SEM, the development of cases-influence diagnostics remains an open area of research. Some case-influence diagnostic methods are based on case-level residual estimates (Bollen & Arminger, 1991; Yuan & Zhong, 2008; Yuan & Hayashi, 2010) and others are based on local influence analysis (Lee & Wang, 1996). With PP-PPMC, we show that the approach as outlined for global model fit assessment can be adapted to automate the generation of  $p$ -values and plots useful for case-influence diagnosis for any appropriately defined test quantity.

### 6.1 Test Quantities for Case-influence Diagnosis

**Classification of Influential Cases** In the context of regression modeling, influential cases can be classified into two categories. Specifically, cases with extreme predictor values are referred to as *leverage points*, while cases with large residuals (i.e. sitting far from the regression line) regardless of predictor values are referred to as *outliers*. When a leverage point has a large (or small) residual, the case is called a bad (good) leverage point (Rousseeuw & Leroy, 1987) in the sense that they lead to decrease (increase) in model fit. Notice that these definitions of outliers and leverage points depends on the

model. An outlier in one regression model may turn out to be a good leverage point in another regression model.

In applications of regression models it is well known that outliers and bad leverage points have disastrous effects on parameter estimation and model fit testing, whereas good leverage points lead to a more accurate regression coefficient estimates and improvement in fit. Diagnostic measures for detecting outliers and leverage points are well developed and routinely available in the output of commercial software for regression analysis (Belsley, Kuh, & Welsch, 1980; Cook, 1986).

In the context of factor analysis models, Yuan and Zhong (2008) showed that similar classifications of influential cases is possible when the factor analysis model is regarded as a multivariate regression model with the factor scores playing the role of the predictors and the observed variables playing the role of responses. Outliers and leverage points can then be defined using estimates of factor scores and residuals. In other words, an observation can be referred to as a *leverage point* when the associated factor score estimate is far from the center of the majority of the factor score estimates. And an observation can be called an *outlier* when the associated residual estimate is large, regardless of the value of factor score estimate. A bad leverage point will have large values both in factor score and residual estimates, whereas a good leverage point will have a small value in the residual estimate. As in regression, the definitions of outliers and leverage points are model-based, and thus the status of a case as an outlier or a leverage point could change under two different factor analysis models. The extensions to general structural equation modeling are examined in Yuan and Hayashi (2010).

In this paper, we adopt the same classifications of the influential cases as defined and examined in Yuan and Zhong (2008) and Yuan and Hayashi (2010). The Bartlett formula is used for factor score estimation (see Yuan & Hayashi, 2010, p. 337 for some desirable

properties of the Bartlett factor score estimates):

$$\hat{\eta}_i = \left( \Lambda' \Psi^{-1} \Lambda \right)^{-1} \Lambda' \Psi^{-1} \left( \mathbf{y}_i^{obs} - \boldsymbol{\mu} \right). \quad (22)$$

And test quantities are formulated so as to be sensitive to discrepancies between observed and replicated cases in their factor scores and residuals estimates.

To check the influence of the  $i$ th case  $\mathbf{y}_i^{obs}$  on model fit, it is natural to leave the observation in question out of the analysis and examine the change in model fit based on the remaining data points. This is a well-established method known as the leave-one-out method in cross-validation. In this paper, the sample with the  $i^{th}$  case deleted is denoted as  $\mathbf{Y}_{(i)}^{obs}$ . Let  $\hat{\theta}_{(i)}$  and  $\hat{\mathbf{V}}_{(i)}$  denote the maximum-likelihood parameter estimates and the associated asymptotic covariance matrix, respectively, based on the leave-one-out sample  $\mathbf{Y}_{(i)}^{obs}$ .

The central idea of the case-influence diagnostic method based on PP-PPMC is to measure the divergence between factor score or residual properties associated with the  $i^{th}$  observed case  $\mathbf{y}_i^{obs}$  and the potential replicated case  $\mathbf{y}_i^{rep}$ . Note that the replicated case should be drawn from the posterior predictive distribution formed with the leave-one-out sample  $\mathbf{Y}_{(i)}^{obs}$ . In other words, the parameter posterior is  $p(\boldsymbol{\theta} | \mathbf{Y}_{(i)}^{obs}, H)$  as opposed to  $p(\boldsymbol{\theta} | \mathbf{Y}^{obs}, H)$ . In PP-PPMC,  $p(\boldsymbol{\theta} | \mathbf{Y}_{(i)}^{obs}, H)$  is further approximated with a multivariate normal with mean vector  $\hat{\theta}_{(i)}$  and covariance matrix  $\hat{\mathbf{V}}_{(i)}$ .

**Test Quantities** Now let us introduce some notation useful for formulating and describing the test quantities to be employed for the case-influence diagnostics. Using the leave-one-out ML estimate  $\hat{\theta}_{(i)}$ , we can produce factor scores and residuals for the leave-one-out sample  $\mathbf{Y}_{(i)}^{obs}$ . Let the centroid of the  $N - 1$  factor scores for cases in  $\mathbf{Y}_{(i)}^{obs}$  be denoted  $\bar{\eta}_{(i)}$  and that of the residuals be  $\bar{\varepsilon}_{(i)}$ . Similarly, let the covariance matrix of the  $N - 1$  factor scores be  $\Omega_{(i)}$ , and that of the residuals  $\Xi_{(i)}$ .

Based on some parameter values in  $\boldsymbol{\theta}$ , we can compute factor score and residual

estimates for the  $i$ th case that was left out,  $\mathbf{y}_i^{obs}$ . Let the factor score and residual estimates for  $\mathbf{y}_i^{obs}$  be denoted as  $\hat{\boldsymbol{\eta}}_i^{obs}$  and  $\hat{\boldsymbol{\varepsilon}}_i^{obs}$ , respectively. For the replicate observation  $\mathbf{y}_i^{rep}$ , let the factor score and residual estimates for  $\mathbf{y}_i^{rep}$  be denoted as  $\hat{\boldsymbol{\eta}}_i^{rep}$  and  $\hat{\boldsymbol{\varepsilon}}_i^{rep}$ , respectively.

A natural choice of the leverage test quantity is the Mahalanobis distance of the factor score estimate of the  $i$ th (observed or replicate) case to the centroid of factor scores for the rest of the sample,  $\bar{\boldsymbol{\eta}}_{(i)}$ . Specifically, the *realized* test quantity is defined as

$$T_{\text{leverage}}(\mathbf{y}_i^{obs}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\eta}}_i^{obs} - \bar{\boldsymbol{\eta}}_{(i)})' \boldsymbol{\Omega}_{(i)}^{-1} (\hat{\boldsymbol{\eta}}_i^{obs} - \bar{\boldsymbol{\eta}}_{(i)}). \quad (23)$$

The *predictive* test quantity is

$$T_{\text{leverage}}(\mathbf{y}_i^{rep}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\eta}}_i^{rep} - \bar{\boldsymbol{\eta}}_{(i)})' \boldsymbol{\Omega}_{(i)}^{-1} (\hat{\boldsymbol{\eta}}_i^{rep} - \bar{\boldsymbol{\eta}}_{(i)}).$$

If the vast majority of the *realized* Mahalanobis distances are larger than the corresponding predictive Mahalanobis distances, then the *predictive p-value* for the  $i$ th case is close to 0 and thus,  $\mathbf{y}_i^{obs}$  may be a potential leverage point.

The test quantity for identifying outliers can be defined similarly. Specifically, the *realized* test quantity is defined as

$$T_{\text{outlier}}(\mathbf{y}_i^{obs}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\varepsilon}}_i^{obs} - \bar{\boldsymbol{\varepsilon}}_{(i)})' \boldsymbol{\Xi}_{(i)}^{-1} (\hat{\boldsymbol{\varepsilon}}_i^{obs} - \bar{\boldsymbol{\varepsilon}}_{(i)}). \quad (24)$$

The *predictive* test quantity is

$$T_{\text{outlier}}(\mathbf{y}_i^{rep}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\varepsilon}}_i^{rep} - \bar{\boldsymbol{\varepsilon}}_{(i)})' \boldsymbol{\Xi}_{(i)}^{-1} (\hat{\boldsymbol{\varepsilon}}_i^{rep} - \bar{\boldsymbol{\varepsilon}}_{(i)}).$$

Again, if the vast majority of the *realized* Mahalanobis distances are larger than the corresponding predictive Mahalanobis distances, then the *predictive p-value* for the  $i$ th case is close to 0 and thus,  $\mathbf{y}_i^{obs}$  may be a potential outlier.

## 6.2 Simulation Studies

In this section, we investigate the performance of the PP-PPMC based case-influence diagnostics using simulated data. Two synthetic data sets of size equal to 206 are generated from a CFA model with three correlated factors. Each of the three factors has three indicators. The model has 9 indicator variables and 23 degrees of freedom. Following the method developed by Yuan and Zhong (2008), 2 outliers, 2 good leverage points, and 2 bad leverage points are introduced, and the last 6 cases in the second data set are replaced by those 6 influential cases. No replacement is made for the first data set, and thus neither outliers nor leverage points are expected to exist.

When the true generating model is fitted to each of the two data sets, the asymptotic chi-square test statistics were 23.75 and 37.50, yielding the associated classical  $p$ -values of 0.43 and 0.03, respectively. These results are expected due to the existence of outliers and bad leverage points in the second data set. It is anticipated that those six influential cases in the second data set and any such cases generated at chance levels in either of the two data sets can be flagged by the proposed PP-PPMC case-influence diagnostics.

To determine whether or not  $\mathbf{y}_i^{obs}$  is a leverage point (or outlier), we compute the case-specific PP-PPMC predictive  $p$ -values, denoted by  $p_{pred}^{(i)}$ , by comparing the realized and predictive Mahalanobis distances of the factor score (or residual) estimate for the  $i$ th case to the centroid of the factor score (or residual) estimates with the  $i$ th case removed.

Panel (a) in Figure 2 presents the results for the first data set wherein no influential cases were intentionally included. Each point in the plot represents a pair of the negative logarithm of the case-specific PP-PPMC predictive  $p$ -value for outliers against the negative logarithm of the corresponding case-specific PP-PPMC predictive  $p$ -values for leverage points. Due to the negative log transformation of the PP-PPMC predictive  $p$ -values, larger values on the horizontal (vertical) axis are more likely to be associated with leverage points (outliers). For ease of interpretation, dotted horizontal and vertical reference lines are added to the scatterplot at the value corresponding to the PP-PPMC

predictive  $p$ -value equal to 0.001. Thus, any observations located above the dotted horizontal (vertical) line can be flagged as potential outliers (leverage points). As expected, none of the cases were flagged as potential outliers or bad leverage points by the proposed method. Interestingly, case 174 is identified as a potential good leverage point, and when deleted, the model fit chi-square test statistic actually deteriorated slightly.

### Insert Figure 2 About Here

Panel (b) in Figure 2 presents the same information for the second data set. As shown in the scatterplot, the proposed methods appear to have correctly identified most of the outliers and leverage points. For example, case 203 is identified as a potential good leverage point, and when deleted, the model fit chi-square statistic indeed deteriorated to 38.96 from 37.50. Case 205 is identified as a potential bad leverage point, and when deleted, the model fit chi-square statistic dropped significantly to 24.63 from 37.50. Other influential cases are also correctly identified as good leverage points, e.g. case 204, or outliers, e.g. case 202, or bad leverage points, e.g. case 206. Interestingly, case 201, added to this data set as an outlier, is not flagged as a potential outlier. Instead cases 6 and 159 are flagged as potential outliers.

These results lend initial support to the proposed PP-PPMC based case-influence diagnostics. Considering that there are more than one influential data points, and the proposed method is based on the leave-one-out method, and that standard maximum likelihood estimation may not be robust in the presence of multiple influential cases, the proposed PP-PPMC based method has exhibited surprisingly good performance.

### 6.3 Open-book Closed-book Data

In this section, we apply the proposed case-diagnostic method to the detection of influential cases in the analysis of the OBCB data. Although it is well known (and was seen in the previous section) that the two-factor CFA model fits well, it would be interesting to examine whether or not the existence of any outliers or bad leverage points

can influence the misfit of the model. Recall that the definitions of outliers and leverage points are model-based, implying that the results of the case-influence diagnostics can differ under different models. Thus, it would also be interesting to examine the case-influence diagnostics under alternative models such as a single-factor CFA model, and compare the results with those obtained from the two-factor CFA model.

### Insert Figure 3 About Here

Figure 3 presents similar scatterplots as those shown in Figure 2. Panel (a) reveals that case 81 is a potential good leverage point under the two-factor CFA model, while Panel (b) reveals that the same case is a potential outlier under the single-factor CFA model. The response vector of the case 81 is  $(3, 9, 51, 47, 40)$ , where the first two test scores are below the mean and the last three test scores are above the mean. This pattern of responses is strongly supportive of the two factor hypothesis (the first two tests are open-book) and is correctly identified as a good leverage point under the two-factor CFA model. For precisely the same reason, case 81 is correctly identified as an outlier under the single-factor CFA model. When case 81 is deleted, the overall model fit chi-square statistics actually increased slightly to 2.40 from 2.07 under the two-factor CFA model, while the chi-square statistic dropped significantly to 5.45 from 8.98 under the single-factor CFA model.

## 7 Discussion

Posterior predictive model checking has emerged as a flexible framework for both overall and targeted model-data fit assessment. When fully Bayesian methods are used to fit highly parameterized latent variable models, the output from posterior sampling (with MCMC) makes it straightforward to conduct PPMC. Recognizing the popularity of ML estimation in structural equation modeling, we propose a hybrid approach.

We regard maximum likelihood as a form of large-sample Bayesian estimation procedure and rely on posterior normality (the Bernstein-von-Mises phenomenon) to construct an approximate posterior predictive distribution using byproducts of maximum

likelihood estimation. In this Poor Person's posterior predictive distribution, the exact parameter posterior is replaced by a multivariate normal distribution with mean vector equal to the ML estimate and covariance matrix equal to the inverse of the information matrix.

We demonstrated the flexibility and computational efficiency of the PP-PPMC method with both overall model fit assessment and case-influence analysis as exemplary contexts. We have also studied the relationship between classical  $p$ -values of model fit test statistics, PPMC predictive  $p$ -values, and the PP-PPMC predictive  $p$ -values. Using the Open-Book-Closed-Book data set, we provide an example of the similarity of predictive  $p$ -value estimates using the Bayesian PPMC and the proposed PP-PPMC methods. We establish a coupling relationship, and use it to demonstrate that overall model fit PP-PPMC predictive  $p$ -values can be approximated easily. It amounts to the adoption of a new reference distribution for standard model fit test statistics that respect the degree of uncertainty in estimating unknown model parameters.

There are a number of important limitations to this research. First, the derivations are exclusively based on normal theory linear structural equation modeling. As observed variables depart from normality, the performance of the proposed PP-PPMC method remains unknown. While other distributions could be considered for the choice of the likelihood, such as a thicker-tailed t-distribution, we chose the normal distribution for analytic simplicity and to keep in line with standard procedures in structural equation modeling. Second, the posterior normality is a large-sample approximation that may or may not be appropriate for all model-data combinations. For smaller  $N$ , the posterior is less peaked and less ideally quadratic, again leading to unknown and potential performance issues. Comprehensive comparisons with alternative posterior approximations, e.g., fully Bayesian MCMC sampling or Rubin's (1987) SIR algorithm, should be performed in future research and unfortunately remains out of scope for the current paper, which aims at introducing PP-PPMC. Finally, we have only considered overall fit and

case-influence. There are other aspects of model fit, e.g., residual dependence, that we have glossed over. We have not even discussed potentials of applying this idea to other forms of model fit indices, of which there are many in structural equation modeling. Nevertheless, we believe the PP-PPMC method to be a simple, non-iterative, and flexible alternative to both classical approaches as well as more modern fully Bayesian methods. We hope the initial evidence gathered here can prompt additional research.

## References

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: A bayesian approach. *Psychometrika*, 67, 49–77.

Arminger, G., & Muthén, B. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, 63, 271–300.

Bayarri, M. J., & Berger, J. O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*, 95, 1127–1142.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, NY: Wiley.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology*, 21, 235–262.

Bollen, K. A., & Stine, R. (1993). Bootstrapping goodness of fit measures in structural equation models. In K. A. Bollen & S. Long (Eds.), *Testing structural equation models* (pp. 111–135). Newbury Park, CA: Sage.

Box, G. E. P. (1979). Some problems of statsitics and everyday life. *Journal of American Statistical Association*, 74, 1–4.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society – Series A*, 143, 383–430.

Browne, M. W. (1974). Generalized least squares estimates in the analysis of covariance structures. *South African Statistical Journal*, 8, 1–24.

Browne, M. W. (1984). Asymptotically distribution-free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean and covariance structure models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York:

Plenum.

Cai, L., & Lee, T. (2009). Covariance structure model fit testing under missing data: An application of the supplemented EM algorithm. *Multivariate Behavioral Research*, 44, 281–304.

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, 30, 178–191.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society – Series B*, 48, 1133–169.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883–904.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the problem of sample size: A clarification. *Psychological Bulletin*, 109, 512–519.

Dahl, F. A. (2006). On the conservativeness of posterior predictive *p*-values. *Statistics & Probability Letters*, 76, 1170 – 1174.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York: Chapman & Hall.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70, 320–328.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 145–162). London: Champman & Hall.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In J. Bernardo, J. P. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 147–167). Oxford: Oxford University Press.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, 71, 369–382.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13, 755–779.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 1–19.

Gelman, A. (2007). Comment: Bayesian checking of the second level of hierarchical models. *Statistical Science*, 349–352.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 733–807.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. New York: Chapman & Hall/CRC.

Glas, C., & Meijer, R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, 27, 217–233.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society – Series B*, 29, 83–100.

Hoijsink, H., & Molenaar, I. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, 62, 171–189.

Janssen, R., Tuerlinckx, F., Meulders, M., & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, 25, 285–306.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and

related Bayes' estimates. *University of California Publications in Statistics*, 1, 277–330.

Le Cam, L., & Yang, G. L. (2000). *Asymptotics in statistics: Some basic concepts*. New York: Springer.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. West Sussex, UK: Wiley.

Lee, S.-Y., Song, X.-Y., & Tang, N.-S. (2007). Bayesian methods for analyzing structural equation models with covariates, interaction, and quadratic latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 404–434.

Lee, S.-Y., & Wang, S. J. (1996). Sensitivity analysis of structural equation models. *Psychometrika*, 61, 93–108.

Lee, S.-Y., & Zhu, H. (2000). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, 53, 209–232.

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 18, 663–685.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33, 519–537.

Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 64, 208–232.

MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Commentary: Hopes and cautions in implementing bayesian structural equation modeling. *Psychological Methods*, 17, 340–345.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

Meng, X.-L. (1994). Posterior predictive *p*-values. *Annals of Statistics*, 22, 1142–1160.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17, 313–335.

Muthén, L. K., & Muthén, B. (1998–2012). *Mplus users guide. seventh edition*. Los Angeles, CA: Muthén and Muthén.

Ogasawara, H. (2001). Approximations to the distributions of fit indexes for misspecified structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 8, 556–574.

Palomo, J., Dunson, D. B., & Bollen, K. (2007). Bayesian structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 163–188). New York, NY: Elsevier.

Pena, D., & Tiao, G. C. (1992). Bayesian robustness functions for linear models. In J. Bernardo, J. P. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 365–389). Oxford: Oxford University Press.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of  $p$  values in composite null models. *Journal of the American Statistical Association*, 95, 1143–1156.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics*, 6, 377–401.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 12, 1151–1172.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82, 543–546.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether

to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 11, 424–451.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, 64, 37–52.

Schott, J. R. (2005). *Matrix analysis for statistics*. New York: John Wiley & Sons.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30, 298–321.

Stern, H. S. (2000). Asymptotic distribution of  $p$  values in composite null models: Comment. *Journal of the American Statistical Association*, 95, 1157–1159.

Tollenaar, N., & Mooijaart, A. (2003). Type i errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology*, 56, 271–288.

Yuan, K.-H., & Bentler, P. M. (2007). Structural equation modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 297–358). Amsterdam: North-Holland.

Yuan, K.-H., & Hayashi, K. (2010). Fitting data to model: Structural equation modeling diagnostics using two scatter plots. *Psychological Methods*, 15, 335–351.

Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology*, 38, 329–368.

## Appendix: Theoretical Derivations of the Coupling Effect

To show the result in Equation (20), we will use  $T_{GLS}(\mathbf{Y}, \boldsymbol{\theta})$  presented in (10) as the test quantity because it is easier to manipulate. Recall from Equation (7), the large-sample distribution of  $\mathbf{s}^{rep}$  (vector of replicate sample first and second moments), conditional on  $\boldsymbol{\theta}$  and  $H$  is

$$\sqrt{N}[\mathbf{s}^{rep} - \sigma(\boldsymbol{\theta})] | \boldsymbol{\theta}, H \stackrel{a}{\sim} \mathcal{N}_{p_*}(0, \Gamma[\sigma(\boldsymbol{\theta})]^{-1}). \quad (25)$$

By Cochran's theorem (Cochran, 1934), the quadratic form  $N[\mathbf{s}^{rep} - \sigma(\boldsymbol{\theta})]' \Gamma[\sigma(\boldsymbol{\theta})] [\mathbf{s}^{rep} - \sigma(\boldsymbol{\theta})]$  is approximately chi-square distributed with  $p_*$  degrees of freedom (see also Theorem 10.9 in Schott, 2005). To derive the desired result in Equation (20), we need the following two propositions.

**Proposition 1** Given  $\boldsymbol{\theta}$  and  $H$ , the quadratic form  $N[\mathbf{s}^{rep} - \sigma(\boldsymbol{\theta})]' \Gamma[\sigma(\boldsymbol{\theta})] [\mathbf{s}^{rep} - \sigma(\boldsymbol{\theta})]$  can also be written as  $NT_{GLS}(\mathbf{Y}^{rep}, \boldsymbol{\theta}) | \boldsymbol{\theta}, H$ , and thus we have

$$NT_{GLS}(\mathbf{Y}^{rep}, \boldsymbol{\theta}) | \boldsymbol{\theta}, H \stackrel{a}{\sim} \chi_{p_*}^2. \quad (26)$$

Due to asymptotic posterior normality of  $\boldsymbol{\theta}$ , we have approximately

$$\boldsymbol{\theta} | \mathbf{Y}^{obs}, H \stackrel{a}{\sim} \mathcal{N}_d(\hat{\boldsymbol{\theta}}, \hat{\mathbf{V}}).$$

Using continuous mapping and the multivariate delta method, an approximate posterior distribution of  $\sigma(\boldsymbol{\theta})$  is

$$\sigma(\boldsymbol{\theta}) | \mathbf{Y}^{obs}, H \stackrel{a}{\sim} \mathcal{N}_{p_*}(\sigma(\hat{\boldsymbol{\theta}}), \Delta(\hat{\boldsymbol{\theta}})\hat{\mathbf{V}}\Delta(\hat{\boldsymbol{\theta}})') \quad (27)$$

where  $\Delta(\hat{\boldsymbol{\theta}})$  is the Jacobian matrix in Equation (3) evaluated at  $\hat{\boldsymbol{\theta}}$ . It follows that

$$(\sigma(\boldsymbol{\theta}) - \mathbf{s}^{obs}) - (\sigma(\hat{\boldsymbol{\theta}}) - \mathbf{s}^{obs}) | \mathbf{Y}^{obs}, H \stackrel{a}{\sim} \mathcal{N}_{p_*}(\mathbf{0}, \Delta(\hat{\boldsymbol{\theta}})\hat{\mathbf{V}}\Delta(\hat{\boldsymbol{\theta}})'), \quad (28)$$

where  $\mathbf{s}^{obs}$  is the vector of observed sample first and second moments. From Equation (9), the covariance matrix  $\hat{\mathbf{V}}$  is equal to  $[\Delta(\hat{\theta})'N\Gamma(\mathbf{s}^{obs})\Delta(\hat{\theta})]^{-1}$ . Consequently, the approximate distribution of the quadratic form  $N[(\sigma(\theta) - \mathbf{s}^{obs}) - (\sigma(\hat{\theta}) - \mathbf{s}^{obs})]' \Gamma(\mathbf{s}^{obs})[(\sigma(\theta) - \mathbf{s}^{obs}) - (\sigma(\hat{\theta}) - \mathbf{s}^{obs})]$ , conditional on  $\mathbf{Y}^{obs}$  and  $H$ , is a central chi-square with  $d$  degrees of freedom. Using this result and after some algebra, we find the difference between two quadratic forms  $N(\sigma(\theta) - \mathbf{s}^{obs})' \Gamma(\mathbf{s}^{obs})(\sigma(\theta) - \mathbf{s}^{obs}) - N(\sigma(\hat{\theta}) - \mathbf{s}^{obs})' \Gamma(\mathbf{s}^{obs})(\sigma(\hat{\theta}) - \mathbf{s}^{obs})$  is distributed as central chi-square random variable with  $d$  degrees of freedom, conditional on  $\mathbf{Y}^{obs}$  and  $H$ . But notice that the difference in these two quadratic forms can be equivalently expressed as  $NT_{GLS}(\mathbf{Y}^{obs}, \theta) - T_{GLS}(\mathbf{Y}^{obs}, \hat{\theta}) | \mathbf{Y}^{obs}, H$ .

**Proposition 2** Thus the following result holds:

$$NT_{GLS}(\mathbf{Y}^{obs}, \theta) - NT_{GLS}(\mathbf{Y}^{obs}, \hat{\theta}) | \mathbf{Y}^{obs}, H \stackrel{a}{\sim} \chi_d^2. \quad (29)$$

**Coupling** Using the two propositions stated in (26) and (29), we can show the key result in Equation (20) as follows:

$$\begin{aligned} & \Pr\{T_{GLS}(\mathbf{Y}^{rep}, \theta) > T_{GLS}(\mathbf{Y}^{obs}, \theta) | \mathbf{Y}^{obs}, H\} \\ &= \Pr\{NT_{GLS}(\mathbf{Y}^{rep}, \theta) > NT_{GLS}(\mathbf{Y}^{obs}, \theta) | \mathbf{Y}^{obs}, H\} \\ &= \int_{\mathcal{E}} \int_{\theta} \mathbf{1}\{NT_{GLS}(\mathbf{Y}^{rep}, \theta) > NT_{GLS}(\mathbf{Y}^{obs}, \theta)\} p(\mathbf{Y}^{rep} | \theta, H) p(\theta | \mathbf{Y}^{obs}, H) d\theta d\mathbf{Y}^{rep} \\ &= \int_{\theta} \int_{\mathcal{E}} \mathbf{1}\{NT_{GLS}(\mathbf{Y}^{rep}, \theta) > NT_{GLS}(\mathbf{Y}^{obs}, \theta)\} p(\mathbf{Y}^{rep} | \theta, H) d\mathbf{Y}^{rep} p(\theta | \mathbf{Y}^{obs}, H) d\theta \\ &\approx \int_{\theta} \Pr\{\chi_{p^*}^2 > NT_{GLS}(\mathbf{Y}^{obs}, \theta) | \theta, H\} p(\theta | \mathbf{Y}^{obs}, H) d\theta \\ &= \int_{\theta} \Pr\left\{\chi_{p^*}^2 > NT_{GLS}(\mathbf{Y}^{obs}, \hat{\theta}) + NT_{GLS}(\mathbf{Y}^{obs}, \theta) - NT_{GLS}(\mathbf{Y}^{obs}, \hat{\theta}) \middle| \theta, H\right\} p(\theta | \mathbf{Y}^{obs}, H) d\theta \\ &\approx \Pr\left\{\chi_{p^*}^2 - \chi_d^2 > NT_{GLS}(\mathbf{Y}^{obs}, \hat{\theta})\right\}. \end{aligned} \quad (30)$$

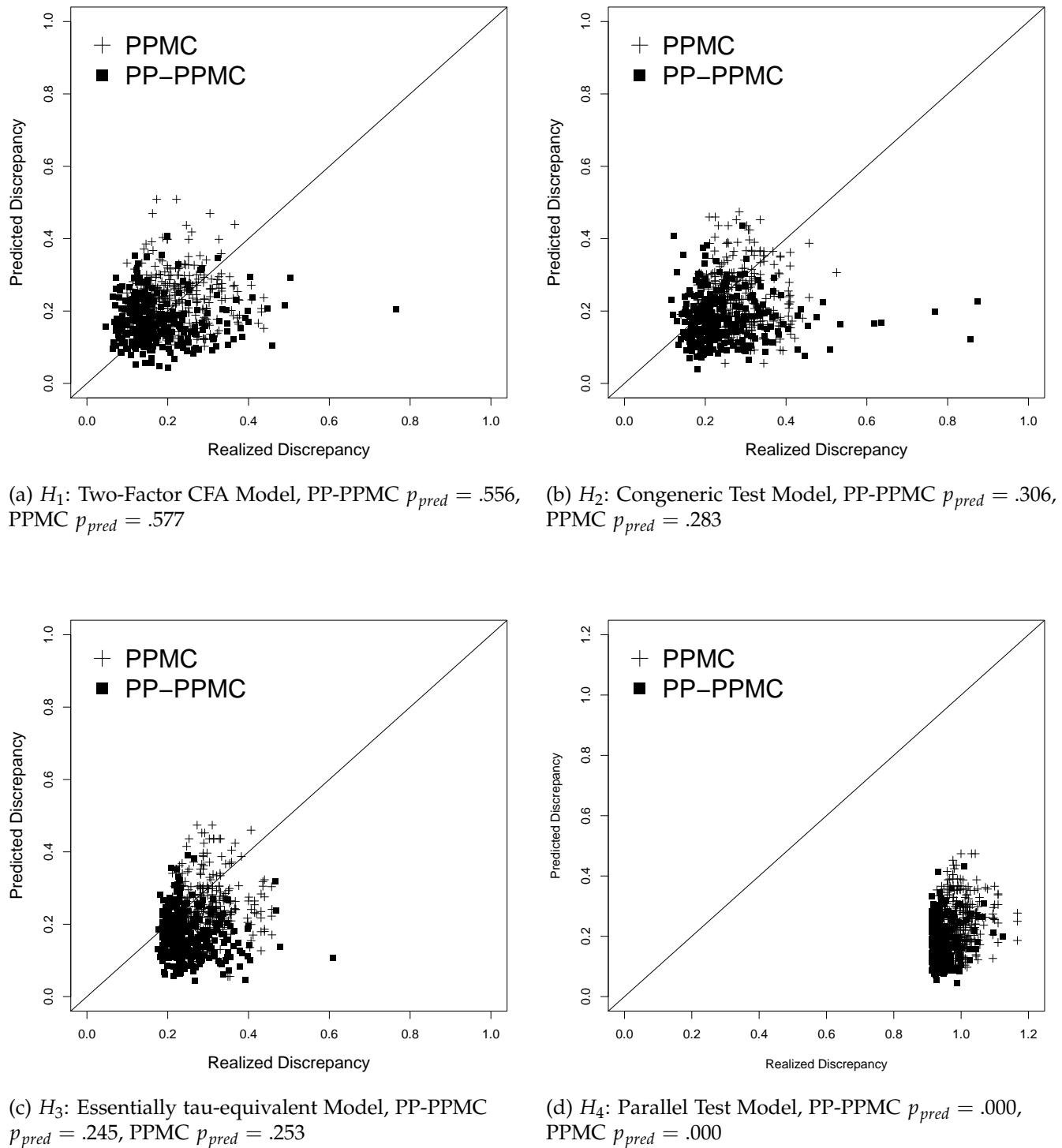


Figure 1: Scatterplots of PP-PPMC and PPMC predictive discrepancies against realized discrepancies for the Open-book Closed-book data set and the four fitted models  $H_1$ ,  $H_2$ ,  $H_3$ , and  $H_4$ .

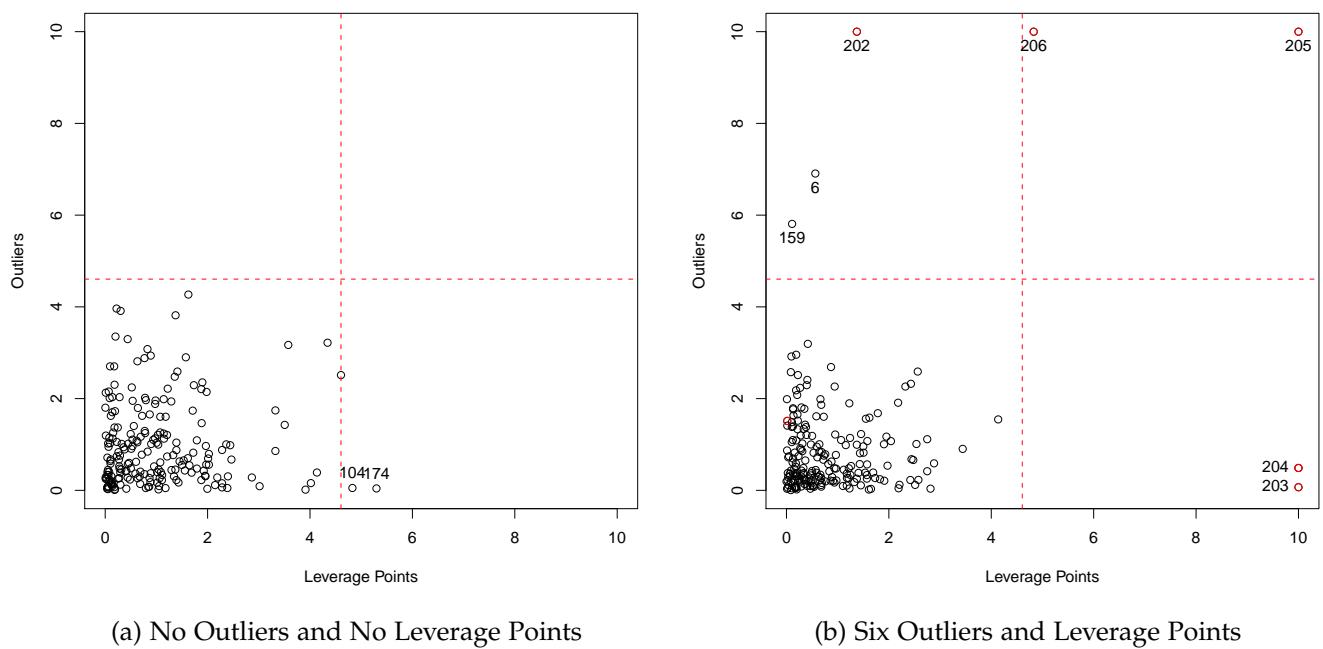


Figure 2: Scatterplots of the Negative Logarithm of the Case-specific PP-PPMC Predictive  $p$ -values for Leverage Points Diagnostics versus the Corresponding Negative Logarithm of the Case-specific PP-PPMC Predictive  $p$ -values for Outliers Diagnosis

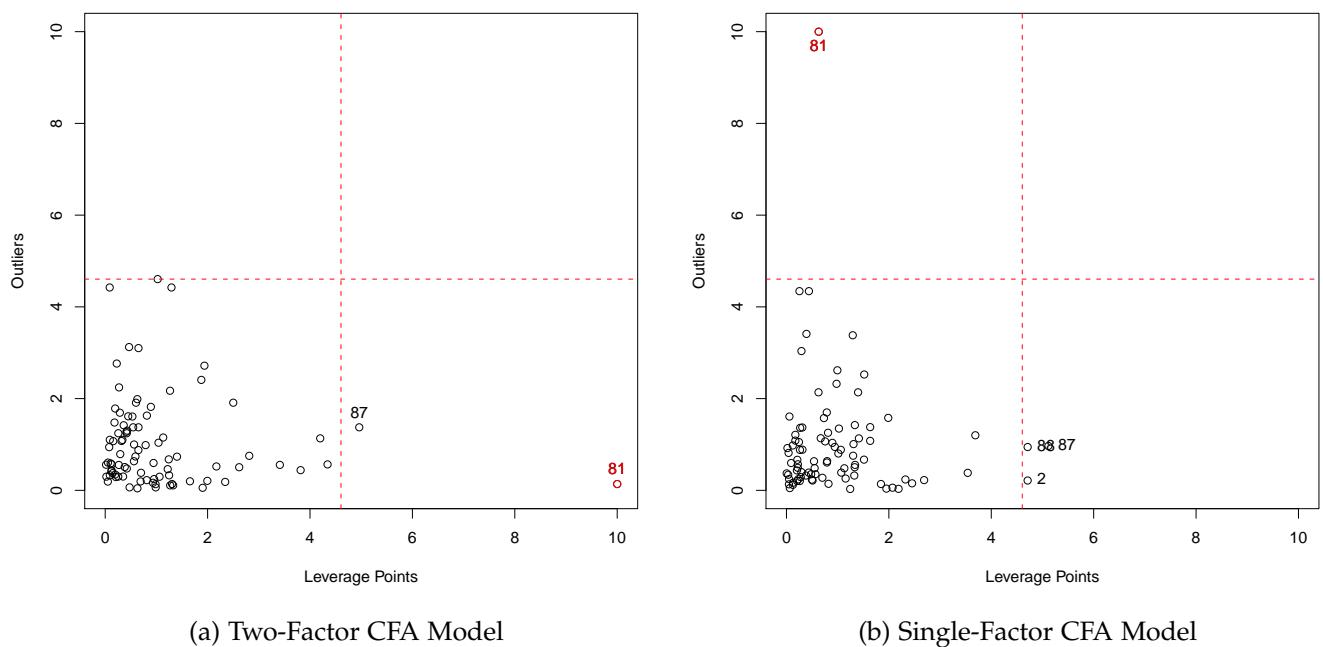


Figure 3: Scatterplots of the Negative Logarithm of the Case-specific PP-PPMC Predictive  $p$ -values for Leverage Points Diagnostics versus the Corresponding Negative Logarithm of the Case-specific PP-PPMC Predictive  $p$ -values for Outliers Diagnosis